Visual Analytics for Comparing the Impact of Outliers in k-Means and k-Medoids Algorithm

Kanika[1], Kanchan Rani[2], Sangeeta, Preeti[3]

[1,2,3]I.K. Gujral Punjab Technical University, Punjab, India
[1]kanikadhanjal@gmail.com, [2]kanchanrana07@gmail.com,
[3]geet.bhagat22@gmail.com, [4]preetynagpal223@gmail.com

Abstract: Clustering is an unsupervised machine learning approach which plays a great role in assigning the data sets into specific clusters based upon some similarity or dissimilarity criterions. K-Means and K-Medoids are the well-known clustering algorithms that are widely used in different application areas of machine learning. K-Means algorithm is sensitive to the outliers due to influence on mean values by outliers in comparison to K-Medoids algorithm which uses medoids, the most centrally located values in a cluster. In this paper, the comparison of both algorithms have been done to evaluate the impact of outliers on their performances by using iris dataset and an interactive web application has been developed with visual analytics to display the impact of outliers on both these clustering algorithms for better insight. The application is accessible through the internet browser.

Keywords: clustering; k-means; k-medoids; outliers; visual analytics.

I. INTRODUCTION

Clustering is grouping of the unlabeled patterns into meaningful clusters. It is one of the most significant approaches of data mining that helps for analysis of data [1]. K-Means and K-Medoids are two well-known clustering algorithms which categorize the data objects into given number of clusters. K-Means clustering algorithm groups the data objects based on the closeness of these objects from each other according to sum of squared Euclidean distances for data objects. On the other hand, K-Medoids is based on the medoids and groups the data objects by minimizing the absolute distance between the data objects and medoids. Thus it is less sensitive to the outliers as compared to the K-Means [2]. An interactive web application has been developed using Shiny web framework for R programming language. Shiny provides the interactive web content by usage of R scripting language [8]. Due to interactive nature of the application, the input values can be changed as per need and the output values get updated immediately to reflect the changes made.

II. RELATED WORK

The detection of outliers and their removal plays an important role for processing meaningful and important data. Zhongxiang Fan proposed an improved K-Means algorithm by detecting outliers based on grid density which reduced the influence of outliers on the results [3].

Surasit Songama et.al proposed a two-phase classification method. In the first phase, the data patterns were clustered by K-Means algorithm and in second phase, outliers were constructed by a distance-based technique and a class label was assigned to each pattern [4].

Fabrizio Angiulli et.al provided distance-based outlier detection method to find out the top outliers in an unlabeled data set [5].

Harshada C. Mandhare performed a comparative study of cluster based outlier detection, distance based outlier detection and density based outlier detection techniques to find out most efficient outlier detection method [6].

III. INPUT DATA AND METHODS

A. Iris Dataset

Dataset used in this research paper is iris obtained from UCI data repository. It contains four number of attributes i.e. Sepal Length, Sepal Width, Petal Length, Petal Width, 150 total number of instances and three number of plant Species i.e. Setosa, Versicolour, Virginica having 50 instances each where each class refers to iris plant.

B. Clustering Algorithm

In this research, two clustering algorithms i.e. K-Means and K-Medoids algorithms have been implemented to perform the clustering on different pairs of attributes of the iris dataset.

C. Outliers Detection

After performing the clustering, then outliers are detected from the dataset having maximum relative distances from the means and medoids of their respective clusters in K-Means and K-Medoids algorithms respectively.

D. Outliers Removal

After detecting the outliers, outliers are removed from the dataset having maximum relative distances from the means

978-1-5386-9346-9/19/$31.00 ©2019 IEEE

and medoids of their respective clusters in K-Means and K-Medoids algorithms respectively.

IV. IMPLEMENTATION

In this research, K-Means and K-Medoids algorithms have been implemented on iris dataset using R language and the visualizations of results have been done on an interactive user interface developed by using Shiny web framework [8] by selecting any two attributes of the iris data set. From this user interface, the user can select any two attributes of iris data, number of clusters to generate (in this case, the number of clusters are three as there are three plant species i.e. Setosa, Versicolour, Virginica), number of outliers to detect and remove and clustering algorithm i.e. either K-Means or K-Medoids algorithm.

The results after clustering are displayed in the form of interactive google bubble chart which is accessible through the internet in which each and every data value can be accessed on moving the cursor to that data point.

The confusion matrix has also been visualized to show that how many instances of a particular Species class have been grouped in a particular cluster and overall accuracy is calculated in terms of the total number of the correct plant species grouped/classified into correct or accurate cluster by the clustering algorithm divided by the total number of instances used for clustering [7].

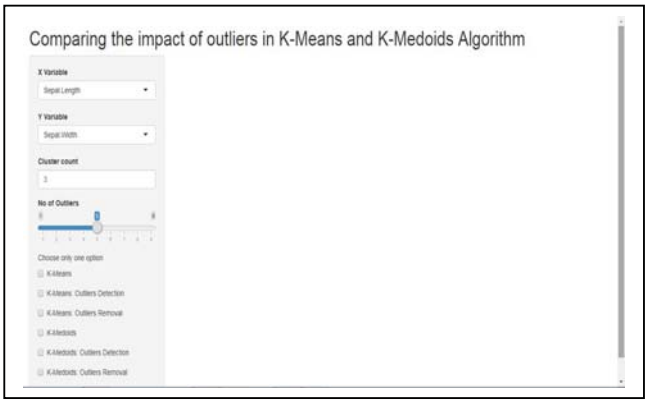

Fig. 1. User Interface developed using the Shiny Package in R.

V. RESULTS AND DISCUSSION

The resulting clusters of K-Means and K-Medoids algorithms are shown in following figures. When we move the cursor on any data point in the graph, the coordinate values of that data point are shown. Hence, the results are more interactive in nature. The results of K-Means algorithm are less stable than the K-Medoids algorithm. Thus, the results of K-Means with highest accuracy have been chosen by running the algorithm multiple times.

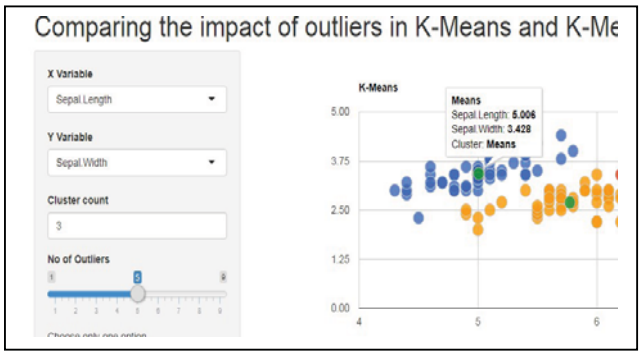

Fig. 2. Clusters created by K-Means algorithm using Sepal Length & Sepal Width and mean values of a cluster shown in green colour.

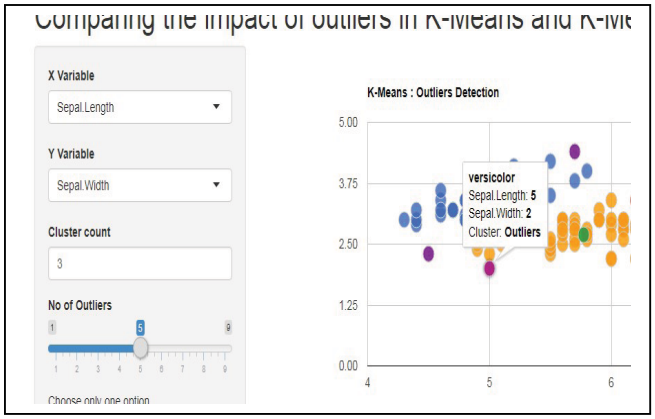

Fig. 3. Clusters created by K-Means algorithm using Sepal Length & Sepal Width and the outliers detected shown in purple colour

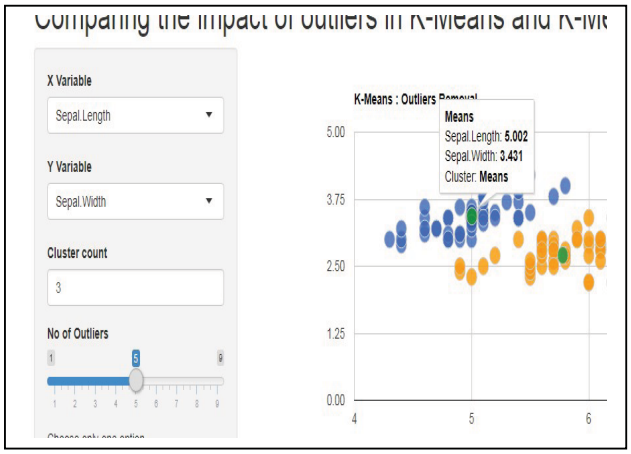

Fig. 4. Clusters created by K-Means algorithm using Sepal Length & Sepal Width and new mean values of a cluster after removal of outliers

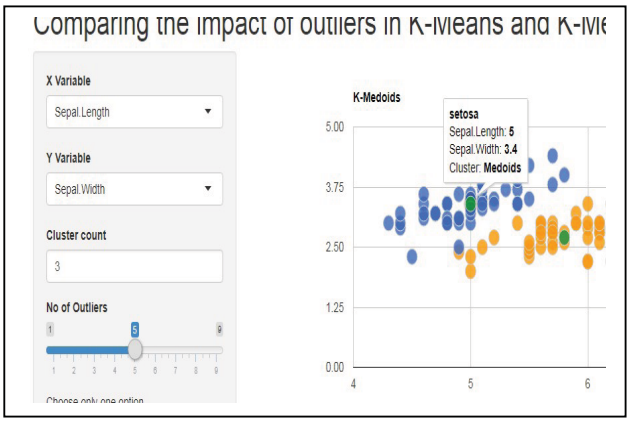

Fig. 5. Clusters created by K-Medoids algorithm using Sepal Length & Sepal Width and medoids of clusters shown in green colour

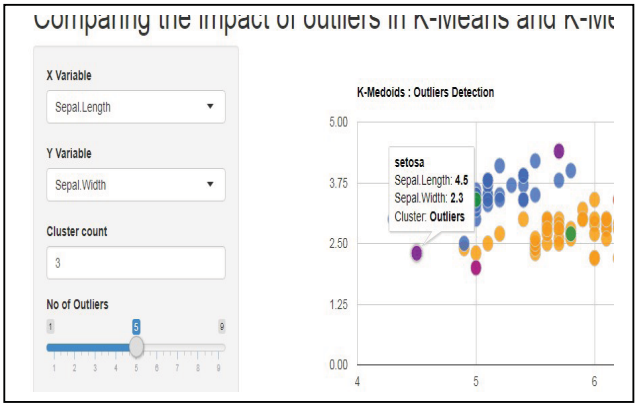

Fig. 6. Clusters created by K-Medoids algorithm using Sepal Length & Sepal Width and the outliers detected shown in purple colour.

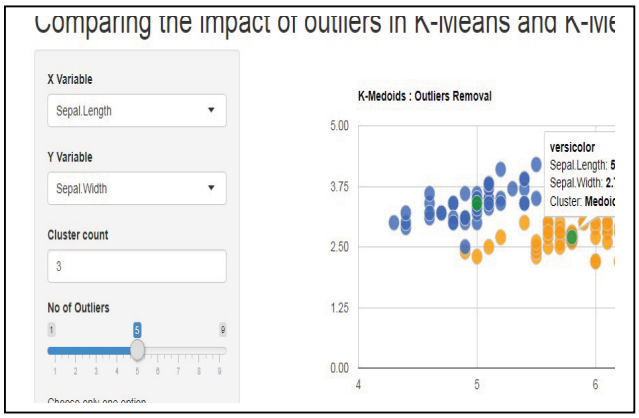

Fig. 7. Clusters created by K-Medoids algorithm using Sepal Length & Sepal Width and same medoids values of a cluster after removal of outliers.

The experiment was performed on the iris dataset by selecting the pairs of any two attributes of the iris dataset to check the accuracy of clustering and outliers' detection and removal for accuracy improvement if possible.

By using the user interface created in web application, we can select any two attributes of the iris dataset at a time to find out if the overall accuracy of K-Means and K-Medoids can be enhanced or not after performing the clustering using two selected attributes and after outliers' detection and removal.

The overall accuracy before and after removing outliers which is being observed in the research has been summarized in the following comparison tables of K-Means and K-Medoids. The entries in tables 'NA' i.e. Not Applicable is for those pair of attributes for which the accuracy doesn't improve after removing the outliers.

TABLE I: CLUSTERING RESULTS USING K-MEANS ALGORITHM

S. No.	Iris Attributes used for clustering	Overall Accuracy (%) of K-Means	No of Outliers removed	Improved Overall Accuracy (%) of K-Means after outliers removal
1	Sepal Length, Sepal Width	82	4	82.19
2	Sepal Length, Petal Length	88	1	88.59
3	Sepal Length, Petal Width	82.66	7	84.61
4	Petal Length, Petal Width	96	NA	NA
5	Sepal Width, Petal Length	92.66	9	92.90
6	Sepal Width, Petal Width	92.66	1	93.28

TABLE II: CLUSTERING RESULTS USING K-MEDOIDS ALGORITHM

S. No.	Iris Attributes used for clustering	Overall Accuracy (%) of K-Medoids	No of Outliers removed	Improved Overall Accuracy (%) of K-Medoids after outliers removal
1	Sepal Length, Sepal Width	82.66	NA	NA
2	Sepal Length, Petal Length	88	1	88.59

3	Sepal Length, Petal Width	82.66	5	85.51
4	Petal Length, Petal Width	94.66	NA	NA
5	Sepal Width, Petal Length	94.66	NA	NA
6	Sepal Width, Petal Width	94	NA	NA

Hence from the comparison tables I and II, it is clear that the accuracy of the K-Means algorithm improves when outliers are removed for five pair of iris dataset attributes out of total six pairs of attributes in comparison to the K-Medoids algorithm whose accuracy improves only for two pair of attributes. Hence, K-Means is more sensitive to the outliers in comparison to the K-Medoids.

Both K-Means and K-Medoids algorithms show highest overall accuracy in clustering the iris data into accurate plant species when the attributes Petal Length and Petal Width are selected. The overall accuracy of K-Means and K-Medoids algorithms in clustering iris data using Petal Length and Petal Width is 96% and 94.66% respectively as shown in figures 8 and 9.

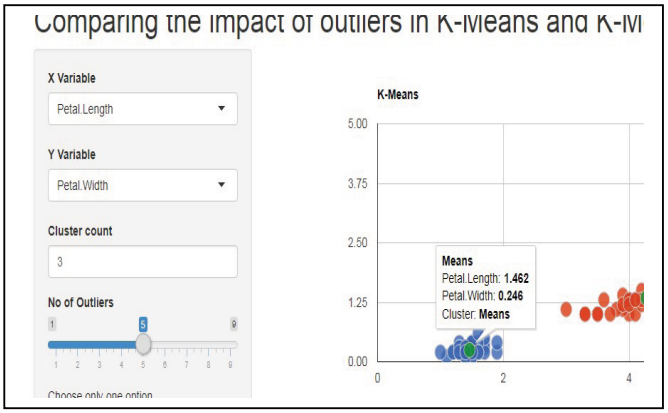

Fig. 8 Clusters created by K-Means algorithm using Petal Length & Petal Width.

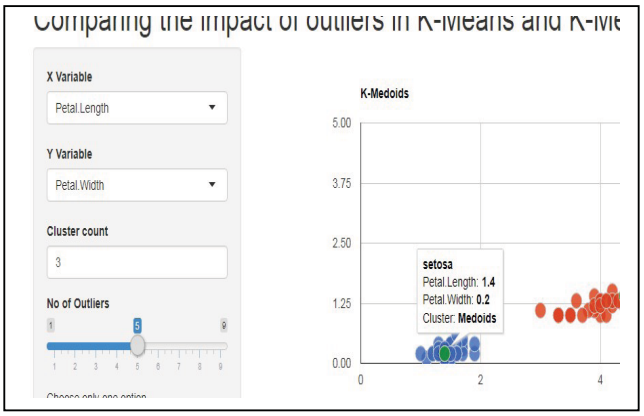

Fig. 9. Clusters created by K-Medoids algorithm using Petal Length & Petal Width

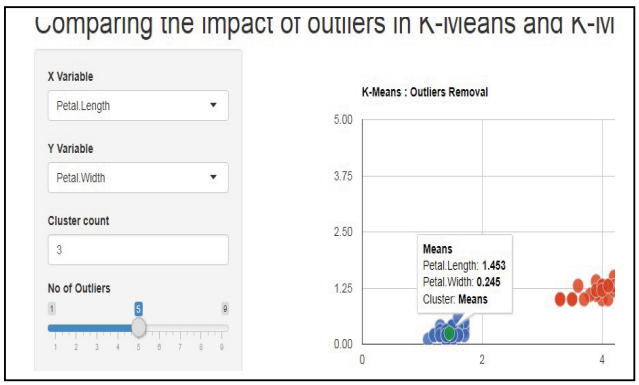

Fig. 10. Clusters created by K-Means algorithm after removal of five outliers using Petal Length & Petal Width

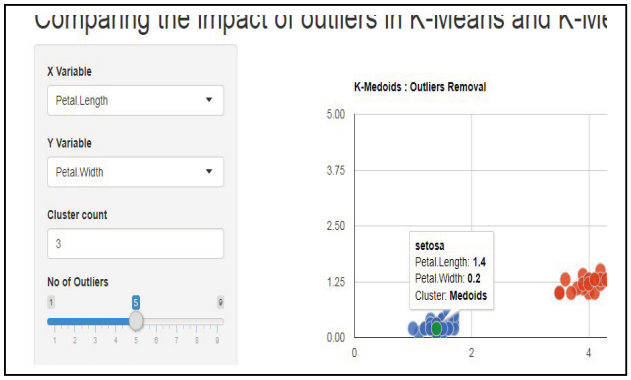

Fig. 11. Clusters created by K-Medoids algorithm after removal of five outliers using Petal Length & Petal Width

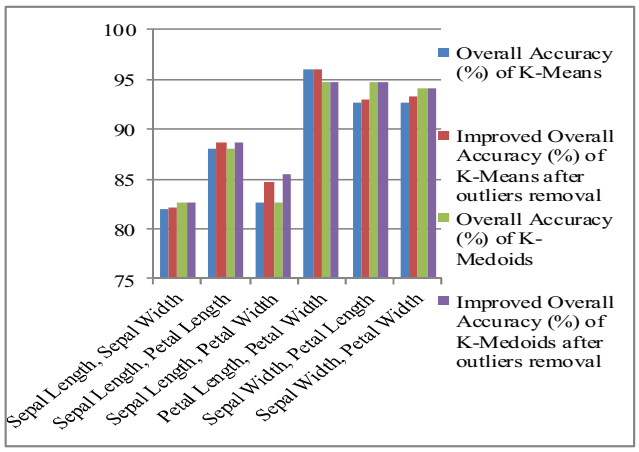

Fig. 12. A graph showing comparison of K-Means and K-Medoids for all pairs of features of iris dataset

However, removing the outliers from the iris data does not improve the overall accuracy of both these algorithms for this pair of attributes i.e. Petal Length and Petal Width of iris dataset as shown in figures 10 and 11.

The above graph shows the comparison of both K-Means and K-Medoids algorithms for all the pairs of attributes of iris dataset. Thus, the pair of Petal Length and Petal Width shows the highest accuracy in classifying the iris data to accurate plant species for both the algorithms.

VI. CONCLUSION

In this research, an interactive web application has been developed using Shiny package and R scripting language which can be accessed through the internet browser. The impact of outliers is more visible to the K-Means algorithm in comparison to the K-Medoids algorithm because medoids are less influenced by the outliers than the means.

Thus, with the help of interactive user interface, it was possible to select the suitable pair of attributes of iris data which have highest accuracy in clustering the iris data into their accurate plant Species. From this research, it has been concluded that the Petal Length and Petal Width attributes play an important role in classifying the iris data into its accurate plant Species. Thus, similar approach can be adopted for other datasets as well to check the impact of dataset attributes and outliers.

REFERENCES

[1] G. Tsoumakas, I. Katakis, I. Vlahavas: Mining multi-label data, in Data mining and knowledge discovery handbook, Springer (2010), pp. 667–685.

[2] R. Capaldo,and F. Collova, “Clustering: A Survey”.

[3] Zhongxiang Fan, San Yun, "Clustring of College Students Based in Improved K-means Algorithm," IEEE International Computer Symposium, 2016.

[4] Surasit Songma, Witcha Chimphlee, Kiattisak Maichalernnukul, Parinya Sanguansat, "Classification via k-Means Clustering and Distance-Based Outlier Detection," IEEE Tenth International Conference on ICT and Knowledge Engineering, 2012.

[5] Fabrizio Angiulli, Stefano Basta, and Clara Pizzuti 2006. "Distance based detection and prediction of outliers," IEEE Transactions on Knowledge and Data Engineering, 2006 18(2), pp. 145-160.

[6] Harshada C. Mandhare. Prof. S.R. Idate, "A comparative Study of Cluster Based Outlier Detection, Distance Based Outlier Detection and Density Based Outlier Detection Techniques," IEEE International Conference on Intelligent Computing and Control Systems, 2017.

[7] Elkan Charles, "Evaluating Classifiers," elkan@cs.ucsd.edu, January 20, 2012.

[8] RStudio and Inc. (2013). Package shiny. https://cran.r-project.org/web/packages/shiny/shiny.pdf